Processing and fusioning multiple heterogeneous information sources in multimodal dialog systems

David Griol, José Manuel Molina, Jesús García-Herrero
Applied Artificial Intelligence Group
Computer Science Department
Carlos III University of Madrid - Spain
Email: {david.griol,josemanuel.molina,jesus.garciaherrero}@uc3m.es

***Abstract*—Context-aware dialog systems must be able to process very heterogeneous information sources and user input modes. In this paper we propose a method to fuse multimodal inputs into a unified representation. This representation allows the dialog manager of the system to find the best interaction strategy and also select the next system response. We show the applicability of our proposal by means of the implementation of a dialog system that considers spoken, tactile, and also information related to the context of the interaction with its users. Context information is related to the detection of user's intention during the dialog and their emotional state (internal context), and the user's location (external context).**

I. Introduction

Multimodal dialog systems [1], [2] are considered as the most appropriate interface for human-computer and human-robot communication in increasingly complex domains, such as Smart Environments[1], health-care [3], or assistance and virtual companions [4], [5]. In such domains, context-awareness plays a very important role, as it is the basis for user-adaptation and proactiveness [6].

Several authors [7], [8] have highlighted the importance of standardizing and sharing a common base for context sensitivity and web services systems. However, most context-aware systems are closed, composed of highly coupled constituents, and generated ad-hoc for a specific domain [8], [9]. The same problem occurs when designing a dialog system. There is a high variety of applications in which dialog systems can be used, some of the most wide-spread are information retrieval from the web [10], database systems [11], and recommendation systems [12].

However, these systems are also usually designed ad-hoc for their specific domain using rule-based models and standards in which developers must specify each one of the steps to be followed by the system. This makes it difficult to adapt the resulting systems to new tasks or incorporate additional context information, as it would require modifying the hand-crafted design, which is very costly in terms of time and effort as this process cannot be automated [13], [14]. In addition, although several works emphasize the importance of taking into account context information not only to solve the tasks presented to the dialog system by the user, but also to enhance the system performance in the communication task, this information is not usually considered when designing a dialog model [15], [16].

[1]http://www.sciencedirect.com/science/article/pii/S0957417411010384

The adaptation capabilities of these interfaces are frequently restricted to static choices made by the users. However, adaptation can play a much more relevant role in speech applications. For example, users have diverse ways of communication. Novice users and experienced users may want the interface to behave completely differently, such as maintaining more guided vs. more flexible dialogs. As stated in [15], processing context is not only useful to adapt the systems' behavior, but also to cope with the ambiguities derived from the use of natural language. For instance, context information can be used to resolve anaphoric references depending on the context of the dialog or the user location. The performance of a dialog system also depends highly on the environmental conditions, such for example whether there are people speaking near the system or the noise generated by other devices.

In order to process context in a meaningful way it is not sufficient to combine input modalities, but also to develop a rich multimodal dialog strategy [17] that allows interpreting the incoming semantic representation of each input modality, evaluating the relevance and completeness of user requests, and identifying and recovering from recognition and understanding errors.

In this paper we propose a framework for the implementation of context-aware multimodal dialog systems that supports the seamless scalability to incorporate new modalities and to adapt to varying applications domains. It is based on a modular architecture that integrates a multimodal fusion model that combines the multimodal information into a single input employed for the selection of the next system action. We also contribute a practical implementation of the method that considers internal and external context in terms of the user's emotional state and the location and temporal context of the interaction respectively.

The remainder of the paper is as follows. Section II describes the proposed framework for the implementation of context-aware multimodal dialog systems. Section III describes the application of our approach to develop a practical system providing academic information. Section IV presents the results of a preliminary evaluation of this practical dialog system. Finally, Section V presents the conclusions and suggests some future work guidelines.

II. Framework for the development of context-aware multimodal dialog systems

A spoken dialog system integrates five main tasks to deal with user's spoken utterances in natural language: automatic speech recognition (ASR), natural language understanding (NLU), dialog management (DM), natural language generation (NLG), and text-to-speech synthesis (TTS). Multimodal dialog systems require additional components respectively related to the fusion and fission the multimodal input and output. This way, the information obtained through the different modalities is managed as a single semantic unit.

We propose to consider an additional module for the integration of the multimodal input to process the context of the interaction. This way, the different modalities are not only integrated to obtain a semantic representation of what the user said, but also to compute the context of the interaction and the user's state, which is essential for the dialog manager to make more informed decisions. Figure 1 shows the proposed framework. As it can be observed, the input to the dialog manager is not only the semantics of the input, but also the external context of the interaction and the internal context that represents information about the user's state at each turn.

Thus, context-awareness is achieved in our proposal with the fusion of two types of context: internal and external. The internal context is based on modeling the user's intentions and emotional states. The external context is based on the physical context of the interaction, mainly on the user location.

A. Internal context

The statistical technique that we propose to model user's intention is described in [18]. A data structure, that we call *User Register* (UR), contains the information provided by the user throughout the previous history of the dialog. For each time i, the proposed model estimates user's intention taking into account the sequence of dialog states that precede time i, the system answer at time i, and the objective of the dialog $\mathcal{O}$. The selection of the most probable user answer U_i is given by:

$$\hat{U}_i = \arg\max_{U_i \in \mathcal{U}} P(U_i | UR_{i-1}, A_i, \mathcal{O})$$

The information contained in UR_i is a summary of the information provided by the user up to time i. That is, the semantic interpretation of the user utterances during the dialog and the information that is contained in a user profile (e.g., user's name, gender, experience, skill level, most frequent objectives, additional information from previous interactions, user's neutral voice, and additional parameters that could be important for the specific domain of the system). We propose to solve the previous equation by means of a classification process, which takes the current state of the dialog (represented by means of the set $UR_{i-1}, A_i, \mathcal{O}$) as input and provides the probabilities of selecting the different user dialog acts.

With respect to the user's emotional state, our emotion recognition method is based on the previous work described in [19], firstly it takes acoustic information into account to distinguish between the emotions which are acoustically more different, and secondly dialog information to disambiguate between those that are more similar. In particular, we discriminate between anger, doubtfulness and boredom.

B. External context

External contextual information is usually measured by hardware or software-based sensors (such as GPS and monitoring programs), or provided by the users. Typically, sensors rely on low level communication protocols to send the collected context information or they are tightly coupled within their context-aware systems. Since sensing techniques are well developed, existing sensors utilize these techniques through instrumentation or polling mechanisms, and extend their capability by acquiring context information from existing systems.

As described in [20], we propose the use of a Facilitator and Positioning Systems to acquire and process external contextual information. The Positioning System communicates with the ARUBA positioning system to extract and transmit positioning information to other agents in the system

The Facilitator System is implemented using the Appear IQ commercial platform (AIQ[2]). The platform consists of two main modules: the Appear Context Engine (ACE) and the Appear Client (AC). The ACE is installed in a server, while the ACs are included in the users' devices.

The ACE implements a rules engine, where the domain-specific rules that are defined determine what should be available to whom, and where and when it should be available. These rules are fired by a context-awareness runtime environment, which gathers all known context information about a device and produces a context profile for that device (e.g., physical location, date/time, device type, network IP address, and user language).

The ACE is divided into three modules that collaborate to implement a dynamic management system that allows the administrator to control the capability of each device once they are connected to the wireless network. The Device Management Module provides management tools to deploy control and maintain the set of mobile devices. The Synchronization Module manages the exchange of files between corporate systems and mobile hand-held devices. Finally, the Device Management is continuously provided with updated versions of the configuration files.

III. A case study: the UAH multimodal dialog system

To show the suitability of our model we have built a multimodal dialog system using a statistical dialog manager and representing the information obtained from the different modalities using EMMA (Extensible MultiModal Annotation markup language[3]).

EMMA is focused on annotating single inputs from users, which may be either from a single mode or a composite input combining information from multiple modes, as opposed to information that might have been collected over multiple turns of a dialog. The language provides a set of elements

[2] www.appearnetworks.com

[3] www.w3.org/TR/emma/

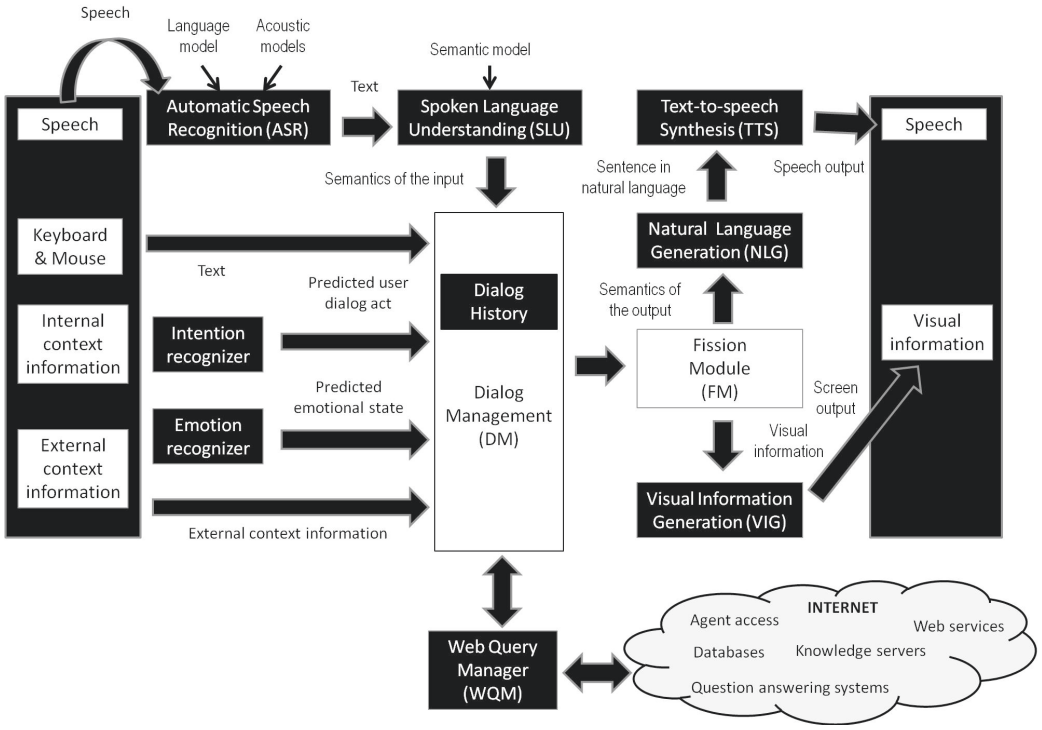

Fig. 1. Proposed framework for the generation of multimodal dialog systems

and attributes that are focused on enabling annotations on user inputs and interpretations of those inputs. The attribute $emma : hook$ can be used to mark the elements in the application semantics within an $emma : interpretation$, which are expected to be integrated with content from input in another mode to yield a complete interpretation.

Following our proposal, the multimodal data fusion and dialog management processes considers the set of input information sources (spoken interaction, visual interaction, user intention modeling, and user emotional state) by means of a machine-learning technique. The dialog manager receives EMMA files containing the results processed by the modules that deal with each input modality. Then, it selects the most appropriate system response using the statistical methodology described in [18], [21] where confidences scores provided by the modules processing each input modality are used in case of conflict among the values provided by several modalities for the same slot. Thus, a single input is generated for the dialog manager to consider the next system response.

Universidad Al Habla (UAH - University on the Line) is a spoken dialog system that provides academic information about the Dept. of Languages and Computer Systems at the University of Granada, Spain. The information that the system provides can be classified in four main groups: subjects, professors, PhD courses and student registration [22].

A corpus of 100 dialogs was acquired with this system from student telephone calls. The total number of user turns was 422 and the recorded speech has a duration of 150 minutes. In order to develop an enhanced version of the system that includes the module shown in Figure 1, we carried out two types of corpus annotation: intentional and emotional.

On the one hand, we estimated the user intention for each user utterance by using concepts and attribute-value pairs. One or more concepts represent the intention of the utterance, and a sequence of attribute-value pairs contains the information about the values provided by the user. We defined four concepts to represent the different queries that the user can perform (*Subject*, *Lecturers*, *Doctoral studies*, and *Registration*), three task-independent concepts (*Affirmation*, *Negation*, and *Not-Understood*), and eight attributes (*Subject-Name*, *Degree*, *Group-Name*, *Subject-Type*, *Lecturer-Name*, *Program-Name*, *Semester*, and *Deadline*). An example of the semantic interpretation of a user's sentence is shown below:

User Turn:
I want to know information about the subject Language Processors I of Computer Science.
Semantic Representation:
(*Subject*)
Subject-Name: Language Processors I
Degree: Computer Science

The labeling of the system turns was similar to that for user turns. To do so, 30 concepts were defined and grouped as task-independent concepts (e.g. *Affirmation* and *Negation*), concepts used to inform the user about the result of a specific query (e.g. *Subject* or *Lecturers*), concepts defined to require the user the attributes that are necessary for a specific query (e.g. *Subject-Name*), and concepts used for the confirmation of concepts and attributes. As shown in Figure 2, the UR defined for the task is a sequence of 16 fields corresponding to the concepts and attributes defined for the task and the user profile.

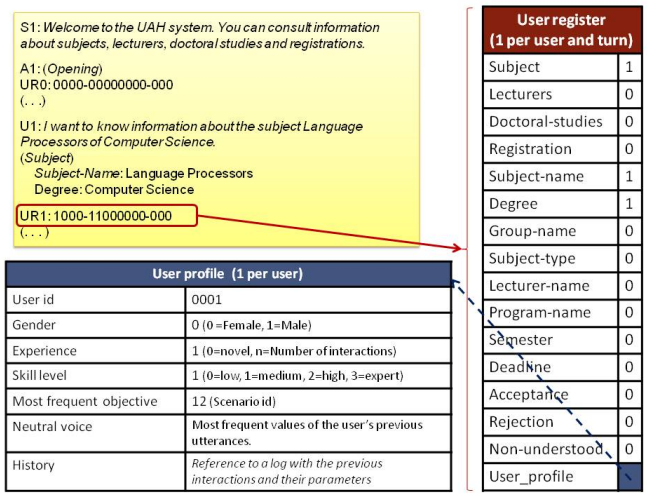

Fig. 2. User Register of the UAH system

On the other hand, we assigned an emotion category (neutral, doubtful, angry, or bored) to each user utterance. Nine annotators tagged the corpus twice and the final emotion for each utterance was assigned by majority voting. A detailed description of the annotation procedure and the intricacies of the calculation of inter-annotator reliability can be found in a previous study [19].

Additionally, we modified the dialog manager to process the user state information in order to reduce the impact of the user negative states and the user experience on the communication, by adapting the system responses considering user states. The dialog manager tailors the next system answer to the user state by changing the help providing mechanisms, the confirmation strategy and the interaction flexibility. The conciliation strategies adopted are, following the constraints defined in [23], straightforward and well delimited in order not to make the user loose the focus on the task.

If the recognized emotion is doubtful and the user has changed his behavior several times during the dialog, the dialog manager changes to a system-directed initiative and generates a help message describing the available options. This approach is also selected when the user profile indicates that the user is non-expert (or if there is no profile for the current user), and when their first utterances are classified as doubtful.

In the case of anger, if the dialog history shows that there have been many errors during the interaction, the system apologizes and switches to DTMF (Dual-Tone Multi-Frequency) mode. If the user is assumed to be angry but the system is not aware of any error, the system's prompt is rephrased with more agreeable phrases and the user is advised that they can ask for help at any time.

In the case of boredom, if there is information available from other interactions of the same user, the system tries to infer from those dialogs what the most likely objective of the user might be. If the detected objective matches the predicted intention, the system takes the information for granted and uses implicit confirmations. For example, if a student always asks for subjects of the same degree, the system can directly disambiguate a subject if it is in several degrees.

In any other case, the emotion is assumed to be neutral, and the next system prompt is decided only on the basis of the user intention and the user profile (i.e., considering user preferences, previous interactions, and expertise level).

IV. Preliminary Evaluation

In order to evaluate our proposal, we have recorded the interactions of 6 recruited users. Four of them recorded 30 dialogs (15 scenarios with the baseline system and 15 with the enhanced system), and two of them recorded 15 dialogs (15 dialogs with the baseline or the enhanced system only). Thus, a total of 150 dialogs were recorded in such a way that there were two dialogs recorded per scenario, three in the case of the five most frequent scenarios of the initial UAH corpus.

TABLE I. Results of the objective evaluation of the systems

Evaluation metrics	Baseline	Enhanced
Dialog success rate	85.0	96.0
Error correction rate	81.0	91.5
Average number of turns per dialog	12.1	8.1
Average number of actions per turn	1.8	1.5
% of different dialogs (intention only)	85.0	83.5
% of different dialogs (intention and emotion)	85.0	88.0
Number of repetitions of the most seen dialog	3.5	6
Number of turns of the most seen dialog	5.5	4.5
Number of turns of the shortest dialog	4.5	4.5
Number of turns of the longest dialog	14.5	12.0

As observed in Table I, on the one hand the success rate for the enhanced system is higher than the baseline. This difference showed a significance of 0.03 in a two-tailed t-test. On the other hand, although the error correction rate is also higher in absolute values in the enhanced system, this improvement is not significant. Both results are explained by the fact that we have not designed a specific strategy to improve the recognition or understanding processes and decrease the error rate. Instead, our proposal for adaptation to the user state overcomes these problems during the dialog once they are produced.

Regarding the number of dialog turns, the enhanced system produced shorter dialogs (with a 0.00 significance value in a two-tailed t-test when compared to the number of turns of the baseline system). As shown in Table I, this general reduction appears also in the case of the longest, shortest and most seen dialogs for the enhanced system. There is also a slight reduction in the number of actions per turn for the dialogs of the enhanced system (with a 0.00 significance value in the t-test). This might be because users have to explicitly provide and confirm more information using the baseline system, whereas the enhanced system automatically adapted the dialog to the user and the dialog history.

Regarding the percentage of different dialogs obtained, the rate was lower using the enhanced system, due to an increment in the variability of ways in which users can provide the different data required to the enhanced system. This result was significant when the dialogs were considered different only when they differed in the sequence of observed user intentions, and also when even with the same sequence of intentions, two dialogs were considered different if the emotions observed were different. This is consistent with the fact that the number of repetitions of the most observed dialogs is higher for the baseline system.

With respect to the dialog participant activity, Figure 3 shows the ratio of user versus system actions. The dialogs of the enhanced system have a higher proportion of system actions due to a reduction of the confirmation turns.

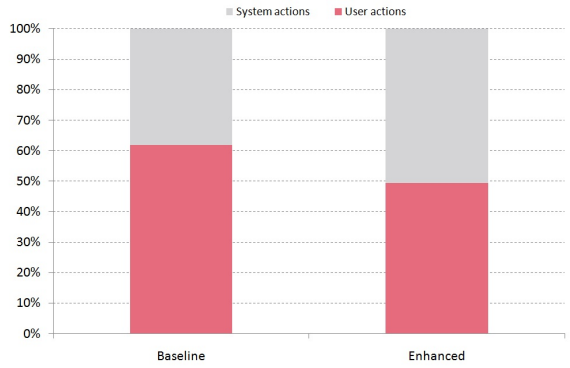

Fig. 3. Ratio of user vs. system actions in the enhanced and baseline systems

Regarding dialog style and cooperativeness, Figures 4 and 5 respectively show the frequency of the most dominant user and system dialog acts in the dialogs collected with the enhanced and baseline systems. On the one hand, Figure 4 shows that users need to provide less information explicitly using the enhanced system, which explains the higher proportion of queries (significant over 98%). On the other hand, Figure 5 shows that there is a reduction in the system requests when the enhanced system is used. This explains a higher proportion of system turns to provide information in the enhanced system.

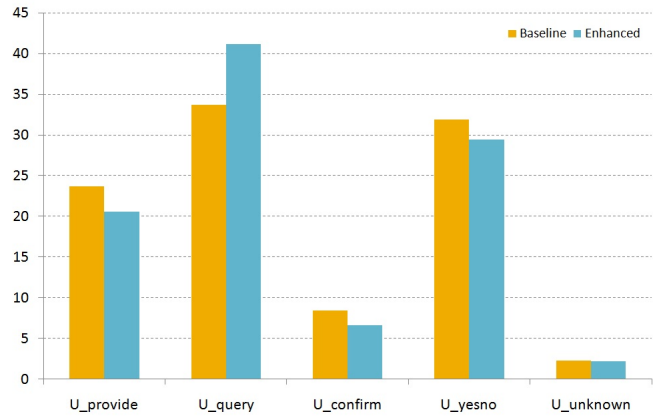

Fig. 4. Histogram of user dialog acts in the enhanced and baseline systems

Table II shows the average results obtained with respect to the subjective evaluation. As can be observed, both systems correctly understand the different user queries and obtain a similar evaluation regarding the user observed easiness in correcting errors made by the ASR module. However, the enhanced system is judged to be better regarding the user observed easiness in obtaining the data required to fulfill the complete set of objectives defined in the scenario, as well as the suitability of the interaction rate during the dialog.

V. Conclusions and Future Work

In this paper we have described a framework to develop multimodal systems that considers information provided by

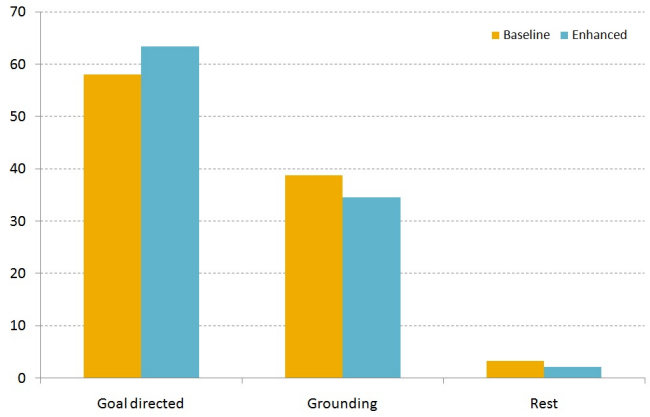

Fig. 5. Histogram of system dialog acts in the enhanced and baseline systems

TABLE II. Results of the subjective evaluation of the systems

Questions (1 to 5 scale)	Baseline	Enhanced
How well did the system understand you?	4.6	4.8
How well did you understand the system messages?	3.6	3.9
Was it easy to obtain the requested information?	3.8	4.3
Was the interaction rate adequate?	3.4	4.2
If the system made errors, was it easy for you to correct them?	3.2	3.3

means of spoken, visual and tactile input modalities. We carry out an additional step towards the adaptation of these systems by also modeling the context of the interaction in terms of external and internal context, which in our case is related to the detection of the user's intention and emotional state.

Several modules have been incorporated in the classical architecture of a spoken dialog system to achieve the integration of the additional input modalities and contextual information sources. These modules respectively allow to predict the next user response for the conversational agent and carry out the fusion of visual and spoken information. The proposed multimodal fusion and dialog management technique allows considering these heterogeneous information sources to select the next system action by means of a classification process. The different methodologies proposed to develop the described modules integrated in the multimodal dialog system have been evaluated in previous works [18], [19], [20], [21].

We have also evaluated the proposed framework with the UAH spoken dialog system, implementing the prediction module between the system's natural language understanding module and dialog manager. Additionally, we have improved the dialog manager to take this information into account in order to compute and adapt the system responses.

The evaluation was carried out using a corpus of interactions of recruited users with the enhanced version of the system. The results show that this version of the system performs better in terms of duration of the dialogs, number of turns needed for successful dialogs, and number of confirmations and repetitions needed. Additionally, the test users judged the system to be better when it could adapt its behavior to their intentions and emotions.

As a future work we plan to annotate the emotions of the

collected corpus in order to refine the adaptation strategies of the dialog manager. We also want to extend the described evaluation with a higher number of users, and also applicate the described framework to develop and evaluate additional practical dialog systems.

ACKNOWLEDGEMENTS

This work was supported in part by Projects MINECO TEC2012-37832-C02-01, CICYT TEC2011-28626-C02-02, CAM CONTEXTS (S2009/TIC-1485).

REFERENCES

[1] R. Pieraccini, *The Voice in the Machine: Building Computers That Understand Speech*. MIT Press, 2012.

[2] W. Wahlster, *SmartKom: Foundations of Multimodal Dialogue Systems Cognitive Technologies*. Springer, 2006, ch. Dialogue Systems Go Multimodal: The SmartKom Experience, pp. 3–27.

[3] I. Bardhan and M. Thouin, "Health information technology and its impact on the quality and cost of healthcare delivery," *Decision Support Systems*, vol. 55, no. 2, pp. 438–449, 2013.

[4] A. Reiss and D. Stricker, "Towards robust activity recognition for everyday life: Methods and evaluation," in *Proc. of 6th Int. Conference on Pervasive Computing Technologies for Healthcare*, 2013.

[5] R. Basole, D. Bodner, and W. Rouse, "Healthcare management through organizational simulation," *Decision Support Systems*, vol. 55, no. 2, pp. 552–563, 2013.

[6] J. Antoniou, C. Christophorou, J. Simoes, and A. Pitsillides, "Adaptive Network-Aided Session Support in Context-Aware Converged Mobile Networks," *Journal of Autonomous and Adaptive Communication Systems*, vol. 5, no. 3, pp. 1–23, 2012.

[7] K. Nihei, "Context Sharing Platform," *NEC Journal of Advanced Technology*, vol. 1, no. 3, pp. 200–204, 2004.

[8] H. L. Truong and S. Dustdar, "A Survey on Context-Aware Web Service Systems," *Journal of Web Information Systems*, vol. 5, no. 1, pp. 5–31, 2009.

[9] B. Brown and R. Randell, "Building a Context Sensitive Telephone: Some Hopes and Pitfalls for Context Sensitive Computing," *Computer Supported Cooperative Work*, vol. 13, pp. 329–345, 2004.

[10] J. Glass, G. Flammia, D. Goodine, M. Phillips, J. Polifroni, S. Sakai, S. Seneff, and V. Zue, "Multilingual spoken-language understanding in the MIT Voyager system," *Speech Communication*, vol. 17, pp. 1–18, 1995.

[11] H. Melin, A. Sandell, and M. Ihse, "CTT-bank: A speech controlled telephone banking system - an initial evaluation," in *TMH Quarterly Progress and Status Report (TMH-QPSR)*, vol. 1, 2001, pp. 1–27.

[12] J. Liu, S. Seneff, and V. Zue, "Dialogue-Oriented Review Summary Generation for Spoken Dialogue Recommendation Systems," in *Proc. of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL*, 2010, pp. 64–72.

[13] T. Paek and R. Pieraccini, "Automating Spoken Dialogue Management Design using Machine Learning: An Industry Perspective," *Speech Communication*, vol. 50, pp. 716–729, 2008.

[14] J. Rouillard, "Web services and speech-based applications around VoiceXML," *Journal of Networks*, vol. 2, no. 1, pp. 27–35, 2007.

[15] S. Seneff, M. Adler, J. Glass, B. Sherry, T. Hazen, C. Wang, and T. Wu, "Exploiting Context Information in Spoken Dialogue Interaction with Mobile Devices," in *Proc. of Int. Workshop on Improved Mobile User Experience*, 2007, pp. 1–11.

[16] J. Ko, F. Murase, T. Mitamura, E. Nyberg, M. Tateishi, and I. Akahori, "Context-Aware Dialog Strategies for Multimodal Mobile Dialog Systems," in *Proc. of AAAI Int. Workshop on Modeling and Retrieval of Context*, 2006, pp. 7–12.

[17] B. Dumas, "Frameworks, description languages and fusion engines for multimodal interactive systems," Master's thesis, University of Fribourg, Fribourg (Switzerland), 2010.

[18] D. Griol, J. Carbó, and J. Molina, "A statistical simulation technique to develop and evaluate conversational agents," *AI Communication*, vol. 26, no. 4, pp. 355–371, 2013.

[19] Z. Callejas and R. López-Cózar, "Influence of contextual information in emotion annotation for spoken dialogue systems," *Speech Communication*, vol. 50, no. 5, pp. 416–433, 2008.

[20] D. Griol, J. Carbó, and J. Molina, "Bringing context-aware access to the web through spoken interaction," *Applied Intelligence*, vol. 38, no. 4, pp. 620–640, 2013.

[21] D. Griol, L. Hurtado, E. Segarra, and E. Sanchis, "A statistical Approach to Spoken Dialog Systems Design and Evaluation," *Speech Communication*, vol. 50, no. 8-9, pp. 666–682, 2008.

[22] Z. Callejas and R. López-Cózar, "Relations between de-facto criteria in the evaluation of a spoken dialogue system," *Speech Communication*, vol. 50, no. 8-9, pp. 646 – 665, 2008.

[23] F. Burkhardt, M. van Ballegooy, K. Engelbrecht, T. Polzehl, and J. Stegmann, "Emotion detection in dialog systems - Usecases, strategies and challenges," in *Proc. of International Conference on Affective Computing and Intelligent Interaction (ACII'09)*, 2009, pp. 1–6.